

# Transferable kriging machine learning models for the multipolar electrostatics of helical deca-alanine

Timothy L. Fletcher<sup>1,2</sup> · Paul L. A. Popelier<sup>1,2</sup>

Received: 19 August 2015 / Accepted: 21 September 2015 / Published online: 17 October 2015  
© The Author(s) 2015. This article is published with open access at Springerlink.com

**Abstract** We exploit the transferability of quantum topological atoms in the construction of a multipolar polarizable protein force field QCTFF. A helical oligopeptide of 10 alanine residues (103 atoms) has its total electrostatic energy predicted using the kriging machine learning method with a mean error of 6.4 kJ mol<sup>-1</sup>. This error is similar to that found in smaller molecules presented in past QCTFF publications. Kriging relates the molecular geometry to atomic multipole moments that describe the ab initio electron density. Atom types are constructed from similar atoms within the helix. As the atoms within a given atom type share a local chemical environment, they can share a kriging model with a reduced number of input descriptors (i.e. features). The feature reduction decreases the kriging training times by more than 23 times but increases the prediction error by only 1.3 %. In transferability tests, transferable models give a 5.7 % error when predicting moments of an atom outside the training set, compared to the 3.9 % error when tested against data belonging to atoms included in the training data. The transferable kriging models successfully predict atomic multipole moments with useful accuracy, opening an avenue to QCTFF modelling of a whole protein.

**Keywords** Multipole moments · QTAIM · Quantum chemical topology · Peptides · Kriging · Machine learning · Alanine · Force field · Electrostatics

## 1 Introduction

The simulation and analysis of large biomolecules is a task best handled by force fields, a state of affairs that will most likely remain so for the foreseeable future due to large computational cost of first-principle methods. The calculation of electrostatic energy deserves special attention given its pivotal role in the modelling of protein structure and dynamics. The electrostatic energy often defies simple computational solutions such as the still-ubiquitous static atom-centred point charges because atomic electron densities are anisotropic and polarise significantly due to changes in molecular geometry. In the past, we have reported on a different and new approach that predicts high-rank multipolar and fully polarised electrostatic interaction energies in water clusters [1], ethanol [2], alanine [3], serine [4], *N*-methylacetamide and histidine [5], aromatic amino acids [6], hydrogen-bonded dimers [7], and atomic kinetic energies of methanol, *N*-methylacetamide, glycine, and triglycine [8]. These studies feature a developing force field, QCTFF, which predicts electrostatic multipole moments using machine learning, fully taking into account polarisation.

The use of multipolar electrostatics is becoming increasingly common, and several groups [9–14] focus on accurately describing molecular and atomic electronic properties without relying on the fitting and parameterisation of point charges. Multipole moments also continue to appear in molecular dynamics simulations [15–18] and geometry optimisations [18–20]. Many current force fields

**Electronic supplementary material** The online version of this article (doi:10.1007/s00214-015-1739-y) contains supplementary material, which is available to authorized users.

✉ Paul L. A. Popelier  
paul.popelier@manchester.ac.uk

<sup>1</sup> Manchester Institute of Biotechnology (MIB), 131 Princess Street, Manchester M1 7DN, UK

<sup>2</sup> School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, UK

compromise for static atom-centred point charges to accumulate computational speed and ease of understanding to end users. However, these point charges typically struggle to reproduce hydrogen bonding and accurate energies in the presence of lone pairs and  $\pi$  systems due to a lack of anisotropy [21]. Additionally, it is recognised that a description of polarisation is a required component [22, 23] of any force field wishing to calculate intermolecular interactions energies [24], and atom-centred point charges cannot capture this anisotropic effect. For modelling biological systems where polarisation effects are deemed vital, polarisation parameters and potentials are often added to existing force fields with additional terms [25, 26]. We instead suggest atom-centred multipole moments (hereon referred to as *moments*) attained through the well-documented topological electron density partitioning scheme [27, 28]. The idea to use the language of dynamical systems and topology [29] to extract insight from any quantum mechanical function is referred to as Quantum Chemical Topology (QCT) [30–32]. Moments taken from the electron density using QCT, intrinsically retain information about the polarisation and charge transfer effects that the atom experiences. Such an increase in sophistication invites an increase in complexity with regard to finding correct potentials and fitting procedures to model the polarisation-sensitive moments, as is done in the force fields XED [33] and AMOEBA [34, 35]. In our approach, we instead suggest the use of machine learning to capture how moments respond to geometric perturbations, thereby capturing both polarisation and charge transfer effects in a unified and streamlined manner [36]. Ultimately these effects express themselves in changes in interatomic electrostatics energies, which are at the heart of structure and dynamics.

In using machine learning models to describe an atom's properties, we introduce a departure from how atom types are approached in typical force fields, particularly with respect to their parameterisation. The popular AMBER [37] force field has led to a traditional atom-type set, GAFF, of 33 basic and 22 special atom types [38], while a method for automated atom typing has been created for use with AMBER [39]. The GAFF parameter set can be used with other force fields such as CHARMM [40] due to an overlap in methodology, but many groups will prefer to create a specialised parameter set to perform a specific function such as peptoid simulation [41], proteins [42], nucleic acids [43], or even molecule-specific moieties [44]. Reparameterisation is a necessary consequence for force fields that neglect polarisation and other directional effects such as hydrogen bonding. Excluding these effects, the modelling of a given atom cannot be consistently accurate in different chemical environments. Even for small molecules such as amino acids, the applicability of parameters must be broken down and new atom types are to be introduced in order to

maintain accuracy. Since parameterisation (and thus atom typing) is often achieved by large-scale fitting of data to parameters, it has been suggested by Maple et al. [45] that high transferability is gained with a high data-to-parameter ratio. Having as few parameters as possible is desirable, so as to keep this ratio high given the large amount of molecular information required for accurate parameterisation. Given a large enough pool of data, automated atom typing procedures [39] are recommended. Another issue is that while many of these parameters claim to be derived from *ab initio* data, the charges are usually the result of fitting to a grid, as is done for RESP [46], which allows charges to violate electronegativity consequences as long as they still reproduce the molecular electrostatic potential.

Through QCT [29], we have access to transferable chemical properties that include polarisation and charge transfer effects, and through moments we can express the electronic properties anisotropically. Original atomic transferability and concomitant atom types were established [31, 47, 48] using cluster analysis operating on atomic properties, such as volume, kinetic energy, charge, and magnitude of the atomic dipole moment. This work reported, for the first time, the *computation* of an atom type. A typical conclusion of this work was that AMBER tends to under-differentiate its carbon atom types but over-differentiate its nitrogen atom types, compared to the QCT atom types. This analysis uses *intrinsic* atomic properties, i.e. the properties do not explicitly express how atoms interact with each other. The next type of transferability analysis [49] used the atomic electrostatic potential as a measure of transferability; this analysis is *extrinsic* in that it invoked a probe that interacts with the atom under study, in this case a proton. The next type of study then embraced the full electrostatic interaction as a transferability gauge, replacing the proton by other topological atoms. This work led to insights into the water trimer and hydrated serine [50] as well as to what extent an atom of interest is influenced by distant atoms in increasingly large chemical environments such as the protein crambin [51].

Proteins are complex materials in that they are constructed from a set of up to 20 building blocks, which are the natural amino acids. Atom typing helps in reducing this complexity to manageable proportions. In fact, oligopeptides already show complex behaviour, compared to the single amino acids they consist of, because of through-space interactions and polarisation effects from neighbouring peptides. A study by Mosquera et al. [52] shows that the electrostatics of a residue in a peptide are affected by the neighbouring residues, and this can therefore be significant if hydrogen bonds are formed or broken. Indeed, as polypeptides tend to fold around themselves, the resulting through-space interactions are amplified in number and strength thus posing a challenge to force fields. Using QCT,

Boyd et al. [53] studied  $\alpha$ -helices and showed that three distinct types of stabilising hydrogen bonds are present.

The development QCTFF needs to embrace this complexity and moves beyond the stage of single molecules and small van der Waals complexes. The ultimate goal of QCTFF is to make predictions for atoms in the condensed matter state. This means the machine learning at the heart of QCTFF needs to train using larger chemical environments. There is a need to address this challenge in a knowledgeable way by detecting and exploiting transferability. In past publications featuring QCTFF, a single machine learning model could only be applied to a single atom that is described in the training data. So far we have worked with small systems that were “closed”, i.e. each atom was aware of the others, and the training was highly specific. In other words, the machine learning model was designed to reproduce exactly what it was trained for, using all information available. Without a method for transferable kriging models, it has not been previously possible to generalise a model to be placed into new systems of interest.

The current work now presents transferable models that can be applied to an atom type, regardless of whether the predicted atom has contributed data to the model's training. For this purpose, we report here on the largest system ever studied using QCTFF: the  $3_{10}$  helix conformation of deca-alanine  $\text{Ala}_{10}$ , consisting of 103 atoms. This system serves as a development ground to answer a number of key questions. Can CPU time be drastically reduced by eliminating non-essential atomic descriptors, within an acceptable threshold of energy error? Can we design a model that can successfully predict properties of many more atoms than it has been trained for? Can a deca-alanine helix be described through atom types by QCTFF resulting in an accurate electrostatic energy prediction? This is the first publication to demonstrate the transferability of QCTFF's multipole moments modelled by kriging.

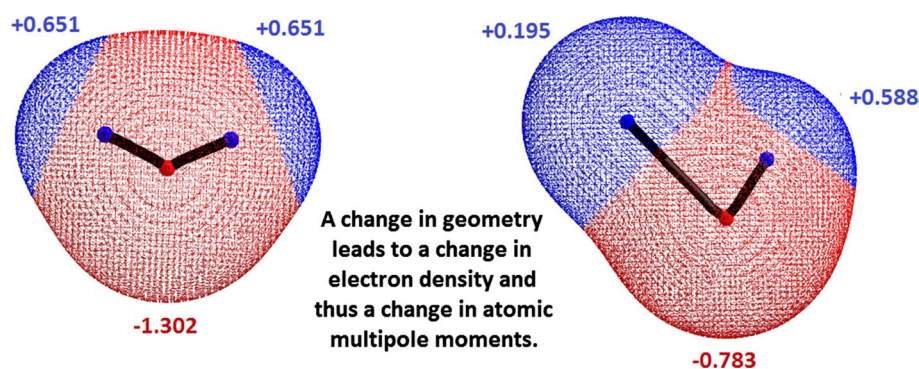
## 2 Methods

A single, helical deca-alanine oligopeptide,  $\text{Ala}_{10}$ , (in particular, a  $3_{10}$ -helix) was extracted from the Brookhaven Protein Data Bank (PDB) (1L36.pdb). Its N terminus peptide bond was broken and the  $(\text{C}=\text{O})\text{C}_\alpha\text{R}$  replaced by a hydrogen. Its C terminus peptide bond was also broken and  $(\text{N}-\text{H})\text{C}_\alpha\text{R}$  replaced by a hydroxyl group. These modifications create a neutral non-zwitterionic system, with a  $\text{COOH}$  and  $\text{NH}_2$  group at either end. The programme Chem3D was used to add hydrogen atoms to the helix. This  $\text{Ala}_{10}$  structure was then geometry optimised at the B3LYP/apc-1 [54] level of theory using the programme GAUSSIAN09 [55] and subsequently distorted by an in-house programme named TYCHE. TYCHE alters a molecule's

geometry through its normal modes of vibration [56] by assigning each (and every) mode a random amount of energy. Bond lengths ( $l_{AB}$ ) were prohibited from stretching beyond the sum of the covalent radii ( $r_A + r_B$ ) of the two bonded atoms  $A$  and  $B$ , determined by a factor  $k_{\text{BOND}}$ . This same factor, which is set to 1.15 in this work, also determined the maximally allowed compression (i.e. 15 % deviation from the reference bond length  $r_A + r_B$ ). More precisely, we have that  $(r_A + r_B)/k_{\text{BOND}} < l_{AB} < k_{\text{BOND}} \times (r_A + r_B)$ . Valence angles  $\alpha$  are also restricted in their motion by a similar window but now determined by the factor  $k_{\text{ANGLE}}$ , which was also set to 1.15. As a result, we have that  $\alpha_0/k_{\text{ANGLE}} < \alpha < k_{\text{ANGLE}} \times \alpha_0$ , where  $\alpha_0$  is a valence angle of the seeding geometry. The final result is a set of 4000 geometrically unique snapshots of the helix, for which GAUSSIAN09 calculates their wavefunctions, again at B3LYP/apc-1 level.

Using default settings, the programme AIMAll [57] partitions space into subspaces that are the topological atoms and integrates over their volume the relevant integrands that correspond to the moments (including monopole and net charge). A topological atom calculated in this way carries an integration error that is output through AIMALL. Should the integration error [often referred to as  $L(\Omega)$  or  $L(\text{Om})$ ] be over the threshold (0.001 au), then this atom's conformation is filtered and not forwarded to become training data, meaning approximately 3000 of the 4000 geometries go on to become training data. The topological atoms [28, 58] do not overlap and do not leave any gaps between them. Moments reflect the shape of the electron density within each topological atom more accurately than a monopole or net charge can. The moments are rotated [59] from the global frame to the atomic local frame (ALF) centred on each atom. The  $x$ -axis of this ALF points from the origin atom to its heaviest bonded neighbour. The  $xy$  plane sweeps out from the  $x$ -axis towards the second heaviest atom bonded to the origin atom. So the origin atom and first and second bonded atoms determine the  $xy$  plane. The  $y$ -axis is then constructed to be orthogonal to the  $x$ -axis and, finally, the  $z$ -axis orthogonal to both, forming a right-handed axis system. The molecular geometry is then converted from Cartesian coordinates in a global frame, to spherical polar coordinates of each atom in the ALF. Note that each atom in the system acts as an origin for its own ALF, allowing the description of the remaining atoms by a unique (but complete) set of spherical polar coordinates. More details can be found in Ref. [60] (including Fig. 1 in that article, showing an example of an ALF).

The atom at the origin sees the rest of the molecule in its own way, that is, from the unique ALF that travels with the atom. This ALF determines the internal coordinates that the atom in question will present to the machine learning method as input, or *features* in machine learning language.



**Fig. 1** Example of the mapping between atomic property (i.e. net atomic charge or output) and the molecular geometry (i.e. input). The constant electron density  $\rho = 0.001$  au contour surface (red = O, blue = H) is shown for two water monomer geometries: (left) the

global energy minimum and (right) a heavily distorted geometry. A change in one O–H bond distance leads to a change in  $\rho$ , to which all moments respond in turn. Here we also show the zeroth moment, or net charge, beside each atom

Each training data point now consists of a set of features that describe the molecular geometry to be matched to that atom's moments. Thus, a change in molecular geometry can be associated with a change in moments. Figure 1 illustrates this association (or input–output mapping) for the water monomer geometry (i.e. input) and the atomic net charges (i.e. output).

The machine learning method kriging [61] (sometimes called ‘Gaussian Process Regression’) is used to model the response of each atom's moments to changes in molecular geometry. These changes occur due to polarisation and manifest themselves in the electron density. Note that the kriging models are trained to reproduce the end point of the polarisation process, not the polarisation itself. Therefore, there is no need for extra terms or corrections focusing on the polarisation energy.

The kriging method is briefly outlined here but has been discussed in greater detail in a previous publication of group [5] and is based on the treatment of Jones et al. [62, 63]. Kriging maps an output's response to any given input. In this case, kriging maps the response of moments to molecular geometries. Equation 1 shows the prediction process

$$\hat{y}(\mathbf{x}^*) = \hat{\mu} + \sum_{i=1}^n a_i \cdot \varphi(\mathbf{x}^* - \mathbf{x}^i) \quad (1)$$

where  $\hat{\mu}$  is the global term, the background value for this output and  $n$  is the number of training geometries. The quantity  $a_i$  is the  $i$ th element of the vector  $\mathbf{a} = \mathbf{R}^{-1}(\mathbf{y} - 1\hat{\mu})$  where  $\mathbf{R}$  is a matrix of error correlations between training points, and  $\mathbf{1}$  is a column vector of ones. The error from the global term is determined by the distance between the new input point ( $\mathbf{x}^*$ ) and a known input point ( $\mathbf{x}^i$ ), each scaled by the magnitude [5]  $\varphi$ . The sum of these errors gives the appropriate deviation from the background term and results

in the new output,  $\hat{y}(\mathbf{x}^*)$ . Thus, for example, the monopole (charge) of an atom can be used as an output that changes in response to a system's geometry. Over a thousand distorted geometries, the monopole moment (“charge” or “Q00”) has a mean ‘background’ value and kriging can map the deviations from the mean as the geometry changes. The fact that kriging uses the distance between the new input and known inputs is chemically sensible as we can assume that if two geometries are very similar, the moments on the atoms in each geometry are similar as well. The symmetric correlation matrix  $\mathbf{R}$  consists of the following kernel,

$$R_{ij} = \exp \left[ - \sum_{h=1}^d \theta_h |x_h^i - x_h^j|^{p_h} \right] \quad (2)$$

where  $d$  is the number of features or the dimensionality of the input space. This value is equal to the number of internal coordinates, which is  $3N - 6$ . The correlation between two points in the training data is thus a function of the distance between the points along with the kriging hyperparameters  $\boldsymbol{\theta}$  and  $\mathbf{p}$ . These two parameters may both be optimised in order for this correlation to best describe the effect that a move between these two inputs has on the selected output. Note that each dimension (feature) of the kriging problem has its own  $\theta_h$  and  $p_h$  value. It has been suggested that  $p_h$  can be fixed at 2 for most cases, and this has held true for past publications, but we continue to optimise this hyperparameter. For that purpose we use particle swarm optimisation [64] against a fitness function described before [5]. This optimisation technique is also used for each  $\theta_h$ .

As the distance between two points closes, their correlation approaches 1 and thus the kriging prediction passes exactly through each training point. Should the new input (for an output to be predicted from) be exactly equal to an existing input, its effect on the output is exactly as it is in



the training set. This means that if we predict the moments for a geometry we have used for training, the kriging reproduces a perfect prediction. This exercise only serves as a “nil test” and is of course not practically useful in terms of real prediction. Conversely, the correlation between two points approaches zero as the distance between them increases, and so it is beneficial to have a high density of training examples in the training space, as this increases the chances of a new input being close to at least one known input and thus having a strong correlation with it. A new input that is found to be outside of the training space (and thus cannot be interpolated) tends towards the global term, and so the model can still attempt a reasonable output when no interpolation is possible.

Each moment on an atom in a new molecular geometry is predicted using kriging models and so each moment will carry an error. This error can be calculated by obtaining the “actual” (i.e. original) ab initio moments for the new geometry and comparing these moments with the predicted moments. Similarly, the predicted moments of two atoms can be used to calculate the interatomic electrostatic energy, and this energy will carry an error resulting from the error in predicting the moments. A total of 25 multipole moments exist on each atom that can be named using the spherical harmonic convention  $Q_{lm}$  where  $l$  is the rank of the moment and  $m$  is the component of that rank. The rank can be considered analogous to atomic orbitals where the zeroth rank (monopole) corresponds to an s-orbital, the first rank (dipole) to a p-orbital, the second rank (quadrupole) to a d-orbital, and so on. Much like atomic orbitals, each rank has  $2l + 1$  components, which leads to a total of 25 ( $= 1 + 3 + 5 + 7 + 9$ ) multipole moments for the first four complete ranks of multipole moments (up to hexadecapole). The monopole moment (rank 0) has a single component, denoted Q00. The dipole moment has three components denoted Q10, Q11c and Q11s. Thus, all Q1k components are dipole moments, Q2k quadrupole moments, Q3k octopoles moments and Q4x hexadecapole moments. These moments have calculable interaction energies that are summed when two atoms interact. The total electrostatic energy error of a system is then defined as in Eq. 3,

$$|E_{\text{system}}^{\text{error}}| = |E_{\text{system}}^{\text{orig}} - E_{\text{system}}^{\text{pred}}| = \left| \sum_{AB} E_{AB}^{\text{orig}} - \sum_{AB} E_{AB}^{\text{pred}} \right| \quad (3)$$

The ‘original’ electrostatic interaction energy ( $E_{\text{system}}^{\text{orig}}$ ) is obtained from the interaction energy from the ab initio moments on two atoms  $A$  and  $B$ . Likewise, a ‘predicted’ interaction energy ( $E_{\text{system}}^{\text{pred}}$ ) can be gained from the interaction of moments obtained from kriging model predictions. The total molecular electrostatic interaction ‘system’ energy is given by the summation of pairwise atomic

interaction energies. The difference between the ‘original’ and ‘prediction’ summations gives the total error in electrostatic energy due to the kriging prediction process,  $|E_{\text{system}}^{\text{error}}|$ . Importantly, each geometry used in such a comparison is part of an *external test set* and is thus never seen by the kriging model training process. The process described in this section is almost entirely automated through a bespoke code ‘GAIA’ developed in-house for the fast creation and application of QCTFF. A flow chart of GAIA, reiterating the steps outlined in this section, can be found in Supporting Information (Figure S1).

In order to establish a basis for transferability, we introduce five types of data sets, as explained in Fig. 2. The details of how these data sets are constructed are given in the figure itself. In order to achieve a proof of concept for transferability, the REGULAR, REDUCED, SHARED, MISMATCHED and TRANSFERRED data sets (and resulting kriging models) must be constructed *in that order*. Each data set is used to test and guide the creation of the next, and the results are divided according to this progression.

### 3 Results and discussion

#### 3.1 REGULAR data sets

Initially, it is important to build REGULAR data sets (as used in previous publications) where every moment on every atom in the system has its own kriging model. Each kriging model has full knowledge of the entire molecular geometry, giving 303 features ( $3N - 6$  where  $N$  is 103, the number of atoms in the system). The REGULAR data sets represent the current paradigm for QCTFF as well as the best model quality currently achievable. Predictions from other data sets in later sections can be compared to the REGULAR prediction errors in order to gauge the quality of newly produced kriging models. Thus, a REGULAR data set acts as a standard or reference point.

Figure 3 shows the deca-alanine helix Ala<sub>10</sub> in its <sub>3</sub><sub>10</sub> conformation with all non-hydrogen atoms numerically labelled. There are 3000 different geometries of this molecule available as training data, potentially, but some of these must be reserved for testing purposes. As few as 1000 geometries can be used for training models without considerable loss of accuracy (see Figure S2 in the Supporting Information) for “REGULAR” training sets. The moments predicted by these models were used to calculate the molecular electrostatic interaction energies of 200 helix test geometries, each geometry involving 4698 unique pairwise atomic interactions. It is useful to keep the test set size at a medium: large enough to obtain an accurate mean error while still small enough for easy analysis.

**Fig. 2** Five data sets used in this study: REGULAR, REDUCED, MISMATCHED, SHARED and TRANSFERRED

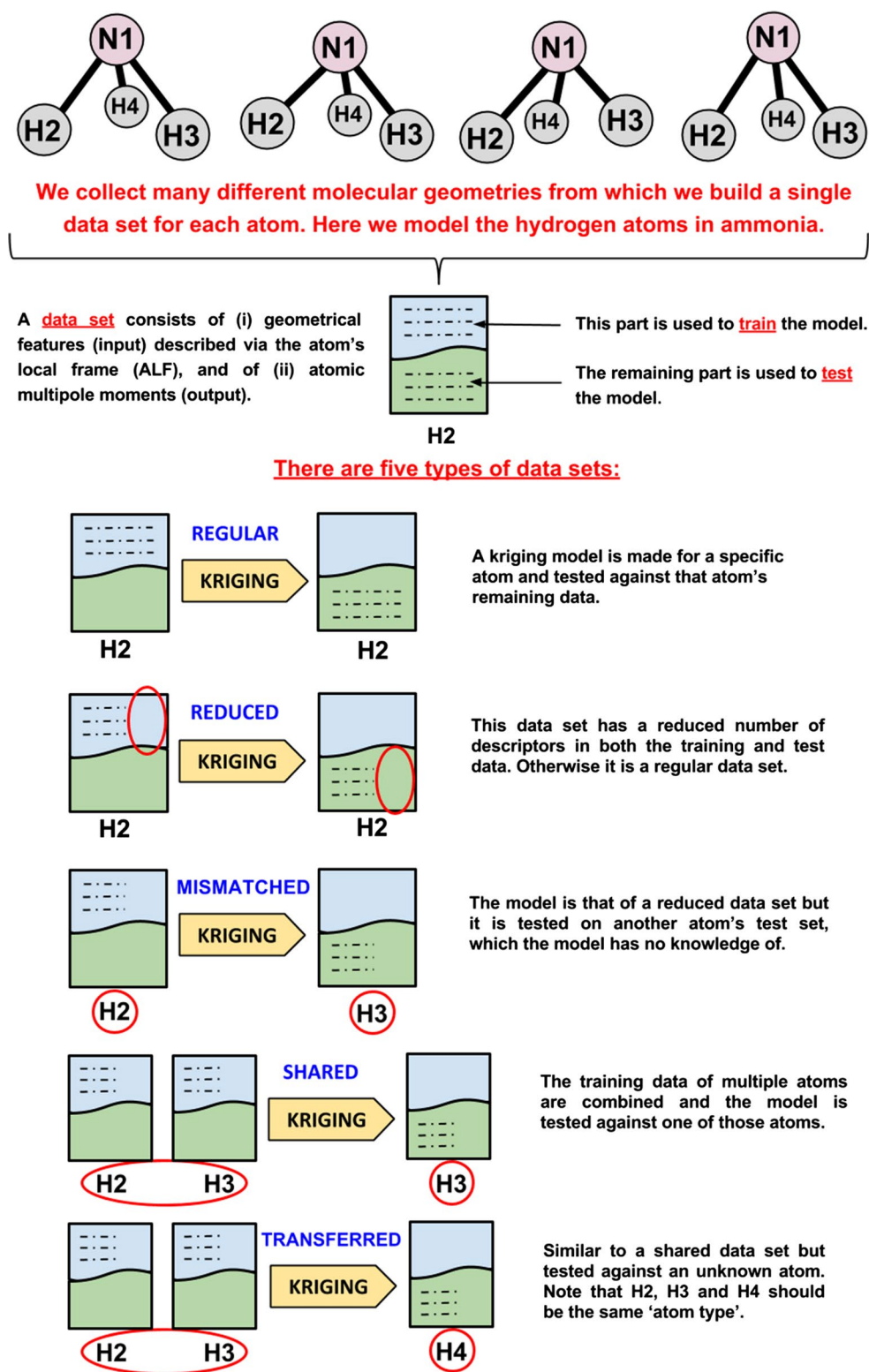
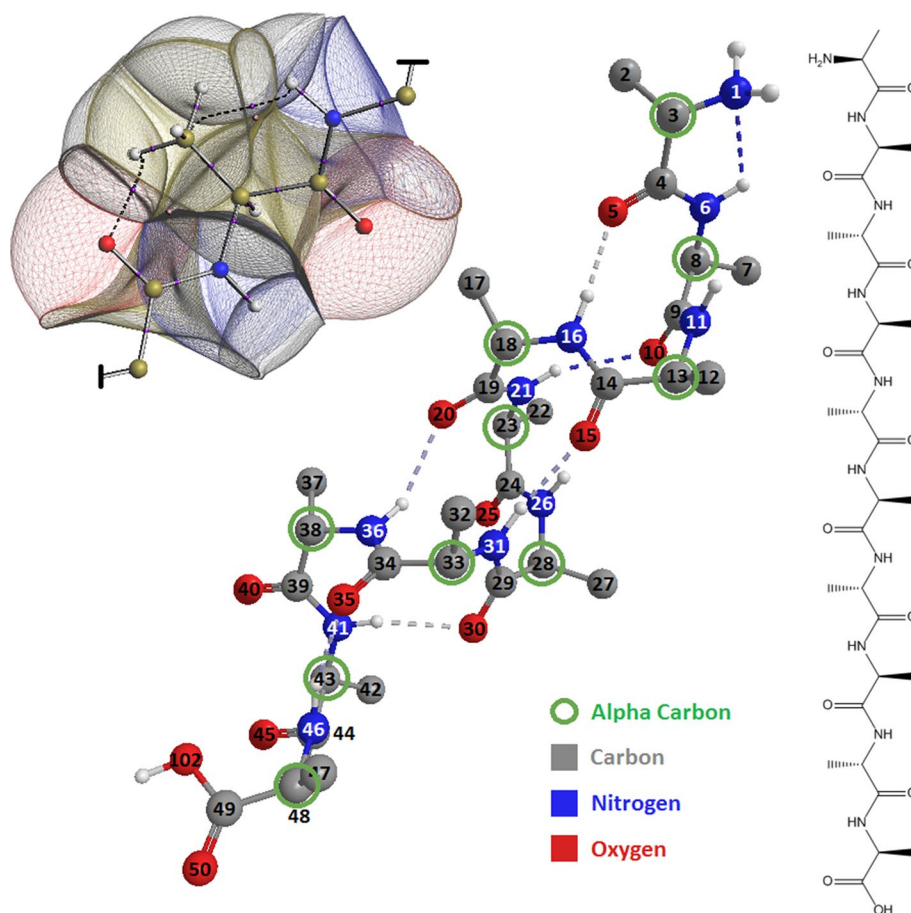


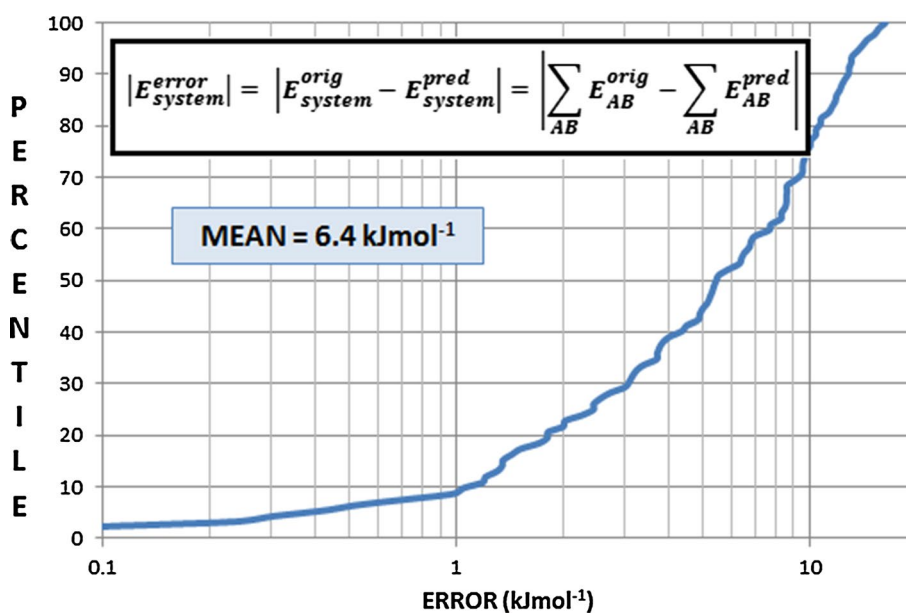
Figure 4 shows a so-called S-curve of prediction errors in the electrostatic interaction energy of the whole Ala<sub>10</sub> helix. More precisely, the S-curve displays the full spectrum of observed  $|E_{\text{system}}^{\text{error}}|$  values (see Eq. 3) for 200 external test geometries.

Each point on the S-curve returns on the  $x$ -axis the error made in predicting the total electrostatic energy of a single geometry in the external test set. The  $y$ -axis consists of a cumulative percentage. As an example of how to read such an S-curve: about 40 % of test geometries have an error

**Fig. 3** Deca-alanine helix. Each alanine unit is chemically (but not geometrically) identical and has an  $\alpha$ -carbon at its centre. Only hydrogen-bonded hydrogens are shown. At the top left is an alanine fragment (with peptide bonds and neighbouring  $C_\alpha$  atoms) magnified as a set of topological atoms superimposed on the molecular graph. The atoms' numerical labels are used throughout the main text



**Fig. 4** S-curve showing the errors in the predictions of the total electrostatic energy,  $|E_{\text{system}}^{\text{error}}|$ , for 200 test geometries of the  $3_{10}$ -helix of deca-alanine



of less than  $4 \text{ kJ mol}^{-1}$  (the oft-quoted  $1 \text{ kcal mol}^{-1}$ ). The bottom left and top right of the S-curve show the best and worst predictions, respectively. The more an S-curve shifts to the left, the better the kriging model predicts the energy.

The  $\text{Ala}_{10}$  helix gives a mean error of  $6.4 \text{ kJ mol}^{-1}$  across the 200 test examples. Although the system is approximately five times the size of a single capped alanine (22 atoms), the mean error is only increased by  $1.1 \text{ kJ mol}^{-1}$ ,

making predictions on realistic biomolecules a feasible goal in the near future. The maximum error (the worst-predicted geometry) of  $16.3 \text{ kJ mol}^{-1}$  is also in line with previously reported doubly capped *single* amino acids.

The number of pairwise interactions increases sharply with the number of atoms in the system. However, cancellations between interaction errors mean that the increased number of interactions does not deteriorate the quality of an S-curve. In other words, only a small increase in error is observed when moving from single amino acids to a chain of 10 amino acids. However, kriging models for individual atoms do not benefit from such cancellations. Hence, studying their accuracy in prediction is a more severe test. Moreover, this study is diagnostically useful and inevitable in the pursuit of insight into a molecular system at atomic resolution. Table 1 makes the latter possible.

We define an average prediction error, which assesses how a kriging model has performed in the prediction of a moment of an atom  $\Omega$ . This error is obtained by averaging this moment  $Q_{lm,\Omega}$  over the values obtained over  $N_{\text{test}}$  test set geometries,

$$\mu_{lm,\Omega} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} |Q_{lm,\Omega,i}^{\text{orig}} - Q_{lm,\Omega,i}^{\text{pred}}| \quad (4)$$

It is fair to express this average relative the test set range of the moment, and conveniently so as a percentage. This error is termed the ‘mean percentage error’ (MPE),

$$\text{MPE}_{lm,\Omega} = 100 \frac{\mu_{lm,\Omega}}{(Q_{lm,\Omega,\text{max}}^{\text{orig}} - Q_{lm,\Omega,\text{min}}^{\text{orig}})} \quad (5)$$

Table 1 lists the errors for the monopole prediction  $Q_{00,\Omega}$  (more precisely the *net* atomic charge actually) on every atom  $\Omega$  in the system. One should keep in mind that Table 1 reports data predicted (by machine learning) rather than the original ab initio data. For example, atom N1 has a mean *prediction* error of 0.001 au for its net charges ranging from  $-1.25$  au to  $-1.37$  au throughout the training set, corresponding to a range of 0.126 au. In this example, the MPE for N1’s net charge is  $<1\%$  of the range. Table 1 is split into 10 sections, each section containing the atoms belonging to a single alanine residue within the oligopeptide chain. Each section (with the exception of the first and last terminal ones) contains a  $C_{\alpha}$  atom (C3 through to C48), the side-chain methyl  $C_{\beta}H_3$  and the nearest peptidic oxygen and hydrogen-substituted nitrogen, NH. No strong patterns exist as to which elements or residues are better *predicted* than others.

The results show that nitrogens, oxygens, carbons and hydrogen have mean MPEs of 1.5, 2.9, 2.3 and 3.3 %, respectively, and all atomic charges throughout all geometries are estimated within 0.3 % of their absolute value.

Although the mean errors are pleasing, the maximum prediction errors are also of concern and their improvement is the subject of ongoing efforts. In the Ala<sub>10</sub> data set, the atom C12 has a significantly higher MPE than other atoms at 8.3 % but is not exceptional in terms of the observed range of feature values or moment value. The next-highest MPE for carbon atoms is three times smaller, at about 2.7 %, of which there are quite a few examples. We feel that the uniquely bad performance of C12 precludes it from being used in non-REGULAR data set (see Fig. 2).

The horizontal dividers in Table 1 separate the Ala<sub>10</sub> helix into 10 single amino acid (alanine) sections. Chemical intuition would suggest that the atoms in each alanine unit in the chain contains approximately equivalent atoms, i.e. that a peptide carbon in one alanine residue is similar to a peptide carbon in another unit, particularly a neighbouring one. Atoms with similar local chemical environments can constitute an atom type if they also share similar chemical properties, in this case (electrostatic multipole) moments.

### 3.2 REDUCED data sets

The route to transferability involves having kriging models that can predict for multiple atoms of a similar type by allowing multiple atoms to share a single kriging model. Before this sharing can occur, it must be decided which atoms share a model (an ‘atom type’). Meanwhile, the features describing these atoms must be reduced to only include a local environment that is common for each of the included atoms.

Data for the (net) atomic charges of all carbon atoms throughout the helix (for every geometry) were collected. Given that each alanine unit has three unique carbon atoms—peptide carbon, alpha carbon, side-chain carbon—we suggest that each of these atoms is equivalent to its counterpart in another alanine unit, thus yielding three carbon atom types. Should these atom types be valid, patterns in their atomic charges can be observed, as is the case in Table 2. In order to establish and test a method for transferability, we have chosen to neglect considering hydrogen atoms on the helix. The hydrogen atoms undoubtedly have ‘types’ but their flexibility and the making or breaking of hydrogen bonds means they are better considered after a proof of concept has been found for other atoms.

It is apparent that atoms of the same type share similar absolute values for their atomic charges and thus have similar ranges across the entire set of data. Minimum values for all side-chain carbons lie between 0.18 and 0.19 au, which is a very tight interval. Similarly, the maximum charges of these side-chain carbons lie within (0.27, 0.28 au), meaning they all share similar ranges and mean values, also. This is a clear-cut example of an atom type occurring in the helix that can potentially be captured by a single kriging model,



**Table 1** Prediction of all atomic charges in the deca-alanine helix and their mean percentage errors (MPE)

N Atom	MPE	O Atom	MPE	C Atom	MPE	H Atom	MPE
N1	1.0	O5	2.3	C2	2.7	H51	2.6
				C3	1.0	H52	3.4
				C4	1.8	H53	1.4
						H54	4.0
						H55	3.1
						H56	2.8
N6	1.3	O10	4.0	C7	2.5	H57	1.9
				C8	1.9	H58	3.2
				C9	2.2	H59	2.9
						H60	3.5
						H61	4.4
N11	1.7	O15	1.2	C12	8.3	H62	2.9
				C13	2.2	H63	3.2
				C14	2.0	H64	2.2
						H65	3.2
						H66	5.7
N16	1.6	O20	3.4	C17	1.4	H67	1.6
				C18	2.3	H68	2.7
				C19	2.1	H69	2.2
						H70	4.3
						H71	5.8
N21	1.5	O25	1.9	C22	2.7	H72	3.9
				C23	1.8	H73	3.7
				C24	2.2	H74	3.6
						H75	2.7
						H76	6.8
N26	1.7	O30	2.3	C27	2.7	H77	2.1
				C28	2.1	H78	1.8
				C29	2.4	H79	3.5
						H80	3.6
						H81	4.5
N31	1.5	O35	2.5	C32	1.4	H82	3.8
				C33	2.0	H83	1.9
				C34	1.7	H84	1.9
						H85	3.9
						H86	4.3
N36	1.7	O40	2.8	C37	2.1	H87	5.0
				C38	2.5	H88	4.2
				C39	1.5	H89	3.3
						H90	3.8
						H91	5.6
N41	1.3	O45	3.7	C42	2.5	H92	2.3
				C43	2.3	H93	1.6
				C44	1.7	H94	1.8
						H95	2.6
						H96	3.8
N46	1.5	O50	5.0	C47	1.3	H97	1.9
		O102	3.3	C48	2.0	H98	2.2
				C49	2.6	H99	2.2
						H100	4.0
						H101	6.1
						H103	4.7

regardless of the residue it belongs to. Peptide and alpha carbons share similar patterns within their own sets, a result that is true for all non-hydrogen atoms in the helix.

However, not all similarities are as high as seen with the side-chain carbons. The lowest value that an alpha carbon's

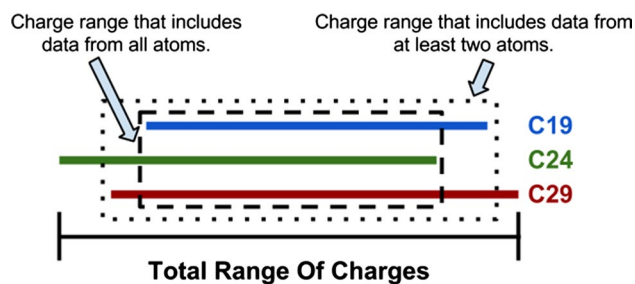
charge adopts lies within the 0.53–0.60 au range, with maximum values lying within 0.62–0.69 au. Although there is now more variation in the observed values, there is considerable overlap between the values for each atom. For example, the values of the charge of the alpha carbon atom C18

**Table 2** Net atomic charge (au) data for each carbon in the deca-alanine helix. There are three classes of carbons: C<sub>α</sub> (alpha), C<sub>β</sub> (side chain) and C(=O) (peptide)

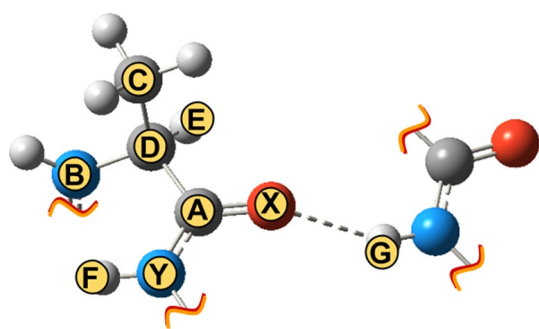
Atom	Min charge	Max charge	Charge range	Mean charge	MPE	Atom type
2	0.19	0.28	0.09	0.23	2.7	Side chain
3	0.55	0.67	0.12	0.60	1.0	Alpha
4	1.71	1.78	0.07	1.75	1.8	Peptide
7	0.19	0.27	0.08	0.23	2.5	Side chain
8	0.54	0.63	0.10	0.58	1.9	Alpha
9	1.76	1.84	0.08	1.80	2.2	Peptide
12	0.19	0.27	0.08	0.23	8.3	Side chain
13	0.56	0.64	0.08	0.59	2.2	Alpha
14	1.75	1.84	0.08	1.79	2.0	Peptide
17	0.18	0.28	0.10	0.23	1.4	Side chain
18	0.53	0.62	0.09	0.58	2.3	Alpha
19	1.76	1.84	0.08	1.80	2.1	Peptide
22	0.18	0.27	0.09	0.23	2.7	Side chain
23	0.54	0.61	0.06	0.58	1.8	Alpha
24	1.77	1.85	0.08	1.81	2.2	Peptide
27	0.18	0.27	0.09	0.22	2.7	Side chain
28	0.56	0.65	0.09	0.60	2.1	Alpha
29	1.74	1.81	0.07	1.77	2.4	Peptide
32	0.19	0.27	0.08	0.23	1.4	Side chain
33	0.59	0.65	0.09	0.60	2.0	Alpha
34	1.77	1.85	0.08	1.81	1.7	Peptide
37	0.18	0.28	0.10	0.24	2.1	Side chain
38	0.54	0.62	0.08	0.58	2.5	Alpha
39	1.78	1.85	0.07	1.81	1.5	Peptide
42	0.19	0.27	0.08	0.24	2.5	Side chain
43	0.55	0.62	0.07	0.58	2.3	Alpha
44	1.75	1.82	0.08	1.79	1.7	Peptide
47	0.18	0.27	0.09	0.23	1.3	Side chain
48	0.60	0.69	0.09	0.65	2.0	Alpha
49	1.80	1.87	0.07	1.84	2.6	Peptide

range from 0.53 to 0.62 au, while alpha carbon C13's range spans the interval (0.56, 0.64 au). We could combine the two data sets to create a range of 0.53–0.64 au, capturing both atoms. This concept is illustrated in the schematic in Fig. 5 where the net charges of three peptide carbons (C19, C24 and C29) are used to create a combined range of data. We expect overlap in the values that these net charges take as the atoms C19, C24 and C29 are all peptide carbons and chemically similar. Indeed, all peptide carbons continue the observed trends. These atoms' charge values largely overlap, their minimum values lying around 1.71–1.80 au and their maxima lying around 1.78–1.87 au.

The question is whether the three atom types suggested in Table 2 can be transferable within the context of kriging machine learning models. Certainly, due to the use of the atomic local frame (ALF), the local environments of atoms



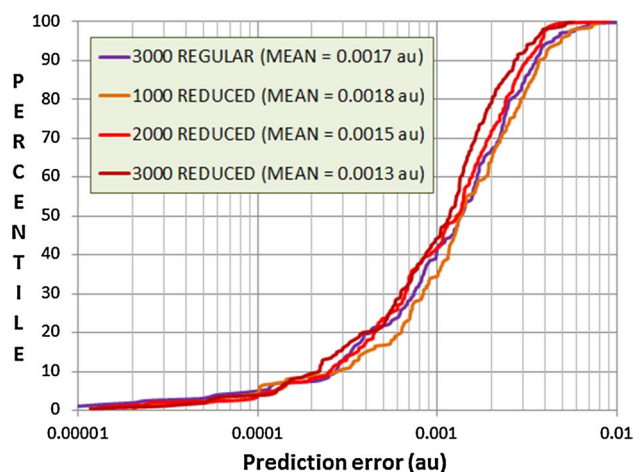
**Fig. 5** Schematic representation for overlap of charge values for three carbon atoms of the same atom type. More overlap is encouraged as it implies the contributing atoms are more similar to one another. Conversely, less overlap is useful as it gives an increased range that this shared model can predict for. Note that atoms are given as examples of atoms that are similar and should have some overlap in their charge values



**Fig. 6** Peptide carbon's (atom A) local environment shown as a cut-out section of the deca-alanine helix. B and Y are peptide nitrogens, C is a side-chain carbon, D is an  $\alpha$ -carbon, while E, F and G are hydrogens. The local environment of atom A includes atoms that define its ALF (atoms X and Y). Other local atoms (B–G) are described by spherical polar coordinates with respect to the ALF on A

can be generalised to transferable units. In a REGULAR data set, an atom's position is described using a full geometric description of the molecule, but this description can be curtailed to a local description, resulting in a REDUCED data set. For example, the first feature in a training set is always the distance between the origin atom (whose kriging model is being made) and its highest priority neighbour, defined by Cahn–Ingold–Prelog priority rules. For peptide carbons, this first feature is the C=O bond distance, regardless of whether this is the C14–O15 bond or the C34–O35 bond. Hence, the training sets of C14 and C34 share this feature, locally. In a helix of 10 repeated alanine residues, some local features in a single alanine residue could be generalised to nine other residues. As an example of this proof of concept, we select four peptide carbons atoms, i.e. C19, C24, C29, C34, (from alanine residues 4, 5, 6, 7 if counting from the N terminus). Note that these atoms were manually selected, but the layout of the data in Table 2 is indicative of how they would be automatically selected in the future. Atoms with similar local chemical environments that share overlapping charge ranges and features could be considered as the same 'type' by an automated procedure. These carbons will share training data of their local geometries. We deliberately selected residues in the middle of the helix in order to avoid edge effects caused by proximity to the helix termini. The local chemical environment and related features for the four peptide carbons are illustrated in Fig. 6.

In order to create a "SHARED" kriging model (that is, one trained with and predicting for multiple atoms, Fig. 2), each atom must be described by the same set of features, applied locally, and one such scheme is suggested in Fig. 6. The reduction in features is accomplished at the discretion



**Fig. 7** S-curves for predictions of the net charge of the  $C_{\alpha}$  carbon C18. REDUCED training sets with different numbers of training examples are compared to the best (largest) training set of the REGULAR kriging models

of the user and is based here on chemical intuition only. We have assumed the need to describe the local non-hydrogen atoms and the nearby hydrogens with considerable net charge, especially those involved in hydrogen bonds. The letter scheme seen in Fig. 5 can be placed on each alanine residue in the helix by replacing the lettered atoms with the numbers of Fig. 3. This local environment is the basis of a REDUCED model, and similar schemes have been applied to other heavy atoms in alanine residues 4, 5, 6 and 7. The result of these schemes are REDUCED training sets and kriging models for peptide carbons,  $\alpha$ -carbons, side-chain carbons, peptide nitrogens and peptide oxygens.

The elimination of features introduces additional prediction error due to the model having a coarser description of the total molecular geometry. Figure 7 shows the representative example of REGULAR and REDUCED models for the net charge (Q00) of atom C18. Surprisingly, a reduction in features leads to a better model as evidenced by lower predictions errors.

It is a common theme in machine learning that the inclusion of too many features can lead to lower prediction accuracy, known as 'over fitting'. In the case of the alpha carbon C18, prediction accuracy is improved (or rather, error is reduced) by 25 % through a reduction in the number of features and increasing the training set size further improves this accuracy. It is not, however, guaranteed that every model is improved through feature reduction, but a close quantitative agreement between REGULAR and REDUCED sets can be assumed given that all errors are below 2 % MPE. It is pleasantly surprising that so little prediction accuracy is sacrificed for a large reduction in model size. Meanwhile, the local environments in the REDUCED

**Table 3** Training times and mean percentage errors (MPEs) of SHARED and REGULAR data sets for the first nine moments of each atom type

Atom type	SHARED atoms	REGULAR CPU training time (min)	REGULAR MPE	SHARED CPU training time (min)	SHARED MPE
Peptide nitrogen	16, 21, 26, 31	1801	4.3	77	5.6
Side-chain carbon	17, 22, 27, 32	1340	1.6	80	2.3
Alpha carbon	18, 23, 28, 33	1605	1.9	33	3.5
Peptide carbon	19, 24, 29, 34	1815	5.5	87	9.2
Peptide oxygen	20, 25, 30, 35	1950	1.8	69	3.3

Each atom type includes four atoms from the helix, numbered according to Fig. 3. For each atom type, the parenthesized atom is the one trained for and predicted in the REGULAR set and the one predicted in the SHARED set

sets can be applied to atoms of the same type throughout the helix, which is the essence of an atom type in QCTFF.

### 3.3 SHARED data sets

We have defined atoms that can share a training set, and we have reduced the number of features to describe only the local geometry. It is now the task to select atoms within the helix to trial a SHARED data set. A SHARED data set should give a kriging model that predicts for all atoms that are used to create the model. Thus, the SHARED data set is more generally applicable than a REGULAR or REDUCED data set. The SHARED kriging models are, however, not yet proof of concept for transferability as they have direct knowledge of the atoms they are attempting to predict.

A SHARED data set can be constructed merely by combining data from many REDUCED data sets for atoms of the same type. For example, C18 and C23 REDUCED data sets can be mixed as they are both  $\alpha$ -carbons and thus have similar local geometries. The resulting SHARED data set can train a SHARED kriging model that can predict for both atoms. In our case, we have determined four atoms to share a data set (and model) for each atom type, listed in Table 3. A key benefit of a SHARED model is in the CPU time saving for training kriging models. The time saving is manifold:

- Only one model needs to be trained for the nitrogen atoms that share the model.
- A reduced number of features means a problem of lower complexity for the kriging machine learning, meaning faster training times.
- A single geometry can be used for many training examples as atoms within geometry can share a kriging model. Hence, fewer geometries are required for the same number of training examples.

This time saving is detailed in Table 3 and shows a dramatic improvement going from REGULAR data sets

to SHARED data sets. Training times can be reduced by approximately  $23\times$  (and even up to  $48\times$  in some cases) at the cost of some prediction accuracy. Each atom type has a single SHARED data set comprised of data for four atoms of the same type and tested against one of those atoms. It is important to note that although the tested atom belongs to the set of trained atoms, the molecular geometries from which the tested atoms originate are not included in the training data. The severity of the incurred error can be interpreted in different ways: the MPE can almost double in some cases, but the original (REGULAR) MPE values are so low that this increase in error does not cause a significant change in prediction accuracy.

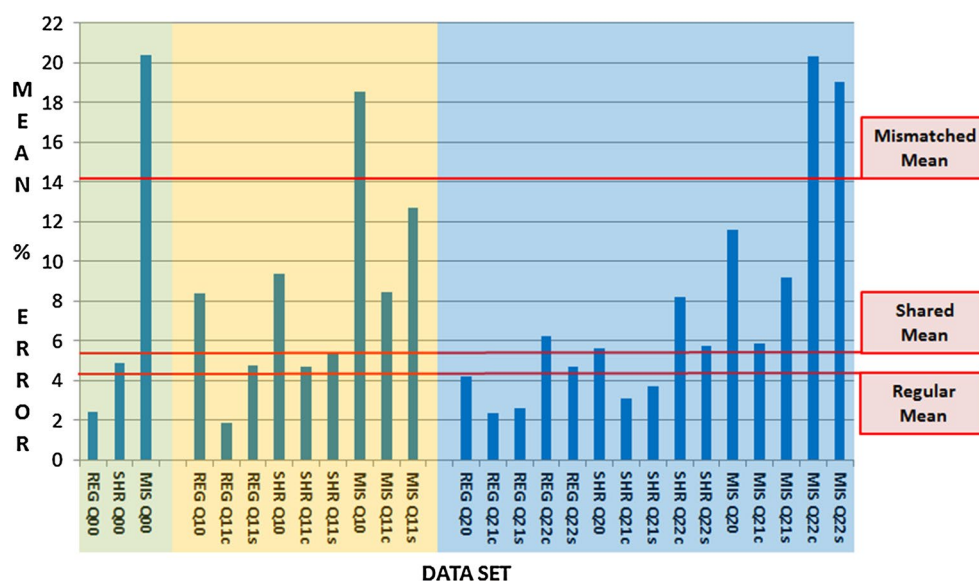
Note that Table 3 includes errors for predictions of monopole, dipole and quadrupole moments for both SHARED and REGULAR data sets. Hence, MPE values are not equal to those found in Table 1, which only included net charges. It is common for the higher rank moments to be more difficult to predict, but they do contribute less to the total electrostatic energy and so their increased MPE is less significant. Furthermore, faster training times allow larger training sets that could help offset this additional error. Although not all listed in Table 3, REDUCED, SHARED and TRANSFERRED sets have identical training times per model due to having an identical number of features. Note that fewer models are required with shared and transferred sets, however. In summary, SHARED models pay for their increased generality with a prediction error penalty. The penalty is small enough that dramatic savings of CPU time make the SHARED models worthwhile. In addition to this advantage, the SHARED models are an important step towards the practical use of transferability.

### 3.4 MISMATCHED data sets

It is important to contrast the SHARED data sets with MISMATCHED data sets. A MISMATCHED data set uses a single atom for prediction and a different atom for testing on. Although the two atoms are of the same type in a MISMATCHED set, no data for the test atom exists in the



**Fig. 8** MPEs for the net charge (green), dipole (yellow) and quadrupole (blue) moments of N21 in the Ala<sub>10</sub> helix. REG (REGULAR) data sets are trained using N21's training set. SHR (SHARED) sets are trained using training sets made up of the data of N21, N26, N31 and N36's with reduced features. MISMATCHED (MIS) sets are trained using N31's REDUCED data set without knowledge of N21



training data. Hence, the test atom should be predicted with higher prediction error than in the SHARED data set, in spite of the SHARED model having increased generality. If a MISMATCHED kriging model has the same predictive power as a SHARED model, then kriging models are inherently transferable with no further effort, and there would be no advantage to SHARED or indeed TRANSFERRED sets (defined in Fig. 2).

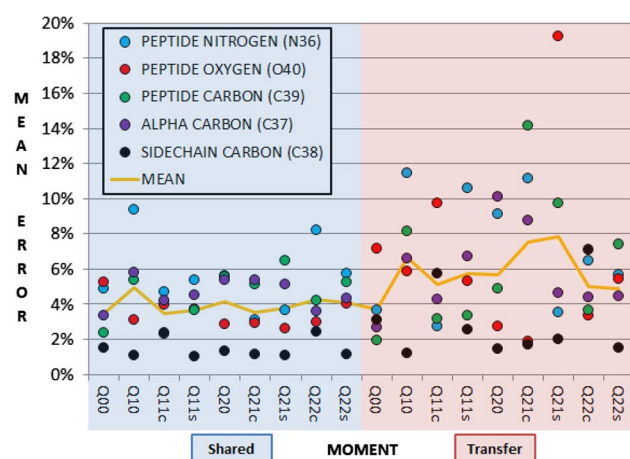
Given that nitrogen atoms normally have moments of high magnitude and are usually difficult to predict, N21 is chosen as an example to demonstrate this. Figure 8 presents the prediction errors of its first nine moments ( $1 + 3 + 5 = 9$ , up to quadrupole moment). Three data sets are compared: REGULAR, SHARED and MISMATCHED. They are most simply contrasted by viewing the red horizontal bars, giving each set's mean error across all moments. The MISMATCHED data training set shows that predictions are generally poor when a training set has no knowledge at all of the atom it is attempting to predict moments for, giving over 20 % error for many moments. Although significantly worse than other prediction errors, the MISMATCHED set still potentially presents a useful prediction. As expected, the REGULAR training set, specific to the atom it predicts for, gives by far the best predictions with most errors ranging between 2 and 4 % MPE. The SHARED training sets (5.6 % mean error) perform more akin to a REGULAR training set (4.3 % mean error) than a MISMATCHED set (14 % mean error).

A common cause of poor predictions can occur when test data lie outside of the training set or when the training set includes some 'bad' data (i.e. the integration error remains stubbornly high). A SHARED training set can offset both of these issues by taking training data from multiple sources, thus allowing for a larger model range and

more training examples. When filtering examples due to high integration error, a larger stock of examples allows for stricter filtering criteria. We can conclude that the SHARED models are much better suited to the prediction of multiple atoms than a MISMATCHED model is. Hence, it is worth constructing a SHARED model if we wish to avoid constructing a kriging model for every atom in a molecule.

### 3.5 TRANSFERRED data sets

The final step towards proof of concept for the transferable kriging models involves the creation of a TRANSFERRED data set by replacing the test data in a SHARED data set with that of another atom. This pooling of data could be seen as analogous to selecting examples from a database in order to build a model suited to your system of interest. The new test atom should be the same atom type that the SHARED data set is composed of. However, the new test atom should also *not* be part of the SHARED training set (see Fig. 2: there is no H4 in the blue training sets). Thus, we are predicting for an atom in an alanine unit that the machine learning has no knowledge of. More precisely, the machine learning has knowledge of alanine units 4, 5, 6, 7 (if counting from the N terminus), and we now use the same models to predict the multipole moments of atoms in the eighth unit. The TRANSFERRED data set could mean the creation of a large 'master' model that contains many atoms (of equivalent type) in the training data and would remain generally applicable through sharing data of many atoms of the same type. This 'master' model could then be applied to other atoms of the same type, using its increased generality to predict atomic multipole moments for which no training data exists. Here, we present five TRANSFERRED data



**Fig. 9** MPEs for all ‘shared’ and ‘transferred’ moment predictions. Two sets of predictions are given, one using SHARED data sets (blue background) and one using TRANSFERRED data sets (red background). An orange line maps the mean error for all predictions of each moment, over all atom types. Atoms predicted in the ‘Transfer’ set are given in brackets the legend

sets. The five SHARED kriging models are tested on an alanine residue within the helix for which the models have no data. The predicted atoms in the TRANSFERRED data sets are atoms N36, C37, C38, C39, O40 (N,  $C_\alpha$ ,  $C_\beta$ , peptide C and O respectively, see Fig. 3), constituting a single alanine residue within the helix (residue 8 if counting from the N terminus). The MPEs for each TRANSFERRED data set are given in Fig. 9 and contrasted with SHARED prediction errors.

It is to be expected that TRANSFERRED data set predictions incur a penalty in accuracy over the SHARED data sets but remain more accurate than a MISMATCHED data set. The MPEs for SHARED data set predictions are low at 3.9 % for all nine moments shown, although they show significant differences between atoms. Peptide nitrogen atoms tend towards being the most difficult to predict and side-chain carbons being the simplest. The TRANSFERRED data set prediction errors tend to be significantly higher than the SHARED data set errors at 5.7 % MPE on average, as was expected. The peptide nitrogen predictions from the SHARED set give 5.6 % mean MPE across all shown moments, which are similar to the TRANSFERRED set’s 7.1 % mean MPE. For the same atom, the MISMATCHED data set gives a prediction of 14 % mean PME. The difference between the TRANSFERRED and MISMATCHED nitrogen prediction errors shows that building a generalised kriging model is better than attempting to directly predict one atom with another atom’s model, even if the two atoms are similar.

In the SHARED data sets, all atom types have similar predictions but in the TRANSFERRED data set there are some clear outliers. The model for moment Q21s of

the peptide oxygen is the most obvious outlier, which has a prediction accuracy significantly worse than all other models. Peptide nitrogen, peptide carbon and peptide oxygen generally present a much more challenging transferability problem than the side-chain carbons and alpha carbons, perhaps due to the large magnitude of their moments. In fact, the side-chain carbon has excellent (>2 % PME) errors in all SHARED and TRANSFERRED sets and represents a highly transferable atom that can be learned from. In spite of fluctuations, most or all of the transferable models built are practically useful and 50 % of the models give a MPE of below 5 %, and are accurate to within 1 % of a moment’s absolute value. Again, we have paid a small penalty in terms of prediction accuracy but added large amounts of functionality to the QCTFF method in the process. TRANSFERRED data sets use working, transferable kriging models that can potentially describe any atom of the same type. These transferable kriging models are a proof of concept for transferability within QCTFF.

## 4 Conclusions

We have presented the first example of transferable kriging models in the QCTFF force field and completed a proof of concept for transferable multipole moment predictions carried out by machine learning. A REGULAR kriging model has full knowledge of the molecular geometry, but by only describing an atom’s local chemical environment, we can reduce a model’s size to a local description of the geometry consequently allowing multiple atoms to share a model. When the SHARED kriging models attempt to predict an atom they have no direct knowledge of, the result is a TRANSFERRED data set. The transferable models require no large alteration to the QCTFF process described in earlier publications and in most cases give close agreement in accuracy to the non-transferable models. Although errors are incurred for both the increased generality of a model (1.3 % mean error above a REGULAR model) and its application to an unknown atom (2.8 % mean error above a REGULAR model), these errors remain low for the majority of transferable models. It is astounding that such excellent models can be made avoiding the typical ‘fine-tuning’ of the entrenched architecture of popular force fields in order to attain quality model accuracy and this leaves much room for improvement, especially for the few models that fail to give good predictions. Only through studying the data sets and determining what makes one models better suited towards transferability than another can we learn to create transferable models for all atoms. Hydrogen atoms have been neglected for the sake of simplicity but should be included in future studies of this type.

The saving of CPU time in moving to REDUCED and SHARED kriging models is highly significant with training times being a factor of 23 shorter in most cases and warrants further investigation even disregarding the goal of transferability.

This proof-of-concept work provides a vital bridge into the future of QCTFF as we move development towards larger biological and condensed phase systems. For the future of QCTFF, REDUCED kriging models means that fewer calculations are required per model and potentially better predictions through a simplified feature space. The SHARED kriging models mean that fewer models need be made as atoms can share a model. When applied to other atoms, the SHARED models (now termed TRANSFERRED models) are actually transferable and allow QCTFF to be applied to a broad range of biological molecules and systems. Large molecules can potentially have accurate multipole moment predictions based on kriging models made using smaller molecules, thus unlocking the potential for extended biological systems to be investigated. Meanwhile, work is continuously underway to deliver faster, more accurate kriging models.

**Acknowledgments** PLA expresses his gratitude to EPSRC for a fellowship (Grant EP/K005472/1), which provides funding for Dr Fletcher.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Handley CM, Hawe GI, Kell DB, Popelier PLA (2009) *Phys Chem Chem Phys* 11:6365
- Mills MJL, Popelier PLA (2011) *Comput Theor Chem* 975:42
- Mills MJL, Popelier PLA (2012) *Theor Chem Acc* 131:1137
- Yuan Y, Mills MJL, Popelier PLA (2014) *J Mol Model* 20:2172
- Kandathil SM, Fletcher TL, Yuan Y, Knowles J, Popelier PLA (1850) *J Comput Chem* 2013:34
- Fletcher T, Davie SJ, Popelier PLA (2014) *J Chem Theory Comput* 10:3708
- Hughes TJ, Kandathil SM, Popelier PLA (2015) *Spectrochim Acta A* 136:32
- Fletcher TL, Kandathil SM, Popelier PLA (2014) *Theor Chem Acc* 133(1499):1
- Gresh N, Cisneros GA, Darden TA, Piquemal J-P (2007) *J Chem Theory Comput* 3:1960
- Price SL, Hamad S, Torrisi A, Karamertzanis PG, Leslie M, Catlow CR (2006) *Mol Sim* 32:985
- Leslie M (2008) *Mol Phys* 106:1567
- Sagui C, Pedersen LG, Darden TA (2004) *J Chem Phys* 120:73
- Ren P, Ponder JW (2002) *J Comput Chem* 23:1497
- Day GM, Motherwell WDS, Jones W (1023) *Cryst Growth Des* 2005:5
- Liem S, Popelier PLA (2003) *J Chem Phys* 119:4560
- Liem SY, Popelier PLA (2008) *J Chem Theory Comput* 4:353
- Liem SY, Popelier PLA, Leslie M (2004) *Int J Quantum Chem* 99:685
- Liem SY, Popelier PLA (2014) *Phys Chem Chem Phys* 16:4122
- Joubert L, Popelier PLA (2002) *Phys Chem Chem Phys* 4:4353
- Shaik MS, Devereux M, Popelier PLA (2008) *Mol Phys* 106:1495
- Cardamone S, Hughes TJ, Popelier PLA (2014) *Phys Chem Chem Phys* 16:10367
- Halgren TA, Damm W (2001) *Curr Opin Struct Biol* 11:236
- Mitin AV (2010) *Int J Quant Chem* 111:2555
- Caldwell JW, Kollman PA (1995) *J Am Chem Soc* 117:4177
- Thole BT (1981) *Chem Phys* 59:341
- Lamoureux G, Roux B (2003) *J Chem Phys* 119:3025
- Bader RFW (1985) *Acc Chem Res* 18:9
- Popelier PLA (2000) *Atoms in molecules. An introduction*. Pearson Education, London
- Popelier PLA (2014) In: Frenking G, Shaik S (eds) *The nature of the chemical bond revisited*. Wiley, Hoboken, Chapter 8, p 271
- Popelier PLA, Brémond ÉAG (2009) *Int J Quant Chem* 109:2542
- Popelier PLA, Aicken FM (2003) *Chem Phys Chem* 4:824
- Popelier PLA (2005) In: Wales DJ (ed) *Structure and bonding. Intermolecular forces and clusters*, vol 115. Springer, Heidelberg, p 1
- Vinter JG (1994) *J Comput Aided Mol Des* 8:653
- Ponder JW, Wu C, Pande VS, Chodera JD, Schnieders MJ, Haque I, Mobley DL, Lambrecht DS, DiStasio RAJ, Head-Gordon M, Clark GNI, Johnson ME, Head-Gordon T (2010) *J Phys Chem B* 114:2549
- Ren PY, Wu CJ, Ponder JW (2011) *J Chem Theory Comput* 7:3143
- Mills MJL, Hawe GI, Handley CM, Popelier PLA (2013) *Phys Chem Chem Phys* 15:18249
- AMBER 9; Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, KMM, Pearlman DA, Crowley M, Walker RC, Zhang W, Wang B, Hayik S, AR, Seabra G, Wong KF, Paesani F, Wu X, Brozell S, Tsui V, Gohlke H, LY, Tan C, Mongan J, Hornak V, Cui G, Beroza P, Mathews DH, Schafmeister C, Ross WS, Kollman PA (2006) *AMBER 9*, University of California, San Francisco
- Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Wang B, Pearlman DA, Crowley M, Brozell S, Tsui V, Gohlke H, Mongan J, Hornak V, Cui G, Beroza P, Schafmeister C, Caldwell JW, Ross WS, AMBER 8, Kollman PA (2004)
- Wang J, Wang W, Kollman PA, Case DAJ (2006) *Mol Graph Model* 25:247
- Vanommeslaeghe K, Hatcher A, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes I, Vorobyov I, McKerell ADJ (2010) *J Comp Chem* 31:671
- Whitelam S, Mirjanián DT, Mannige RV, Zuckermann RN (2014) *J Comput Chem* 35:11
- MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE III, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998) *J Phys Chem B* 102:3586
- MacKerell ADJ, Wiorkiewicz-Kuczera J, Karplus M (1995) *J Am Chem Soc* 117:11946
- Autenreith F, Tajkhorshid E, Baudry J, Luthey-Schulten Z (2004) *J Comput Chem* 25:10

45. Maple JR, Hwang MJ, Stockfish TP, Dinur U, Waldman M, Ewig CS, Halger AT (1993) *J Comput Chem* 15:162
46. Bayly CI, Cieplak P, Cornell WD, Kollman PA (1993) *J Phys Chem* 97:10269
47. Popelier PLA, Aicken FM (2003) *J Am Chem Soc* 125:1284
48. Popelier PLA, Aicken FM (2003) *Chem Eur J* 9:1207
49. Popelier PLA, Devereux M, Rafat M (2004) *Acta Cryst A* 60:427
50. Rafat M, Shaik M, Popelier PLA (2006) *J Phys Chem A* 110:13578
51. Yuan Y, Mills MJL, Popelier PLA (2014) *J Comp Chem* 35:343
52. Lorenzo L, Moa MJG, Mandado M, Mosquera RAJ (2006) *Chem Inf Mod* 46:2056
53. Boyd RJ, LaPointe SM, Farrag S, Boho'riquez HJ (2009) *J Phys Chem B* 113:8
54. Jensen F (2002) *J Chem Phys* 117:9234
55. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JJA, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam NJ, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas Ö, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ, Gaussian, Inc. (2009) Wallingford CT
56. Ochterski JW (1999) Vibrational analysis in Gaussian. [http://www.gaussian.com/g\\_whitepap/vib.htm](http://www.gaussian.com/g_whitepap/vib.htm)
57. Keith TA (2013) AIMAll (Version 13.10.19). <http://aim.tkgrist-mill.com>
58. Bader RFW (1990) *Atoms in molecules. A quantum theory*. Oxford University Press, Oxford
59. Su ZW, Coppens P (1994) *Acta Cryst A* 50:636
60. Mills MJL, Popelier PLA (2014) *J Chem Theory Comput* 10:3840–3856
61. Matheron G (1963) *Econ Geol* 58:21
62. Jones DR, Schonlau M, Welch WJ (1998) *J Global Optim* 13:455
63. Jones DR (2001) *J Global Optim* 21:345
64. Kennedy J, Eberhart RC (1942) *Proc IEEE Int Conf Neural Netw* 1995:4